PATENT ABSTRACTS OF JAPAN

(11)Publication number : 11-184825
(43)Date of publication of application : 09.07.1999

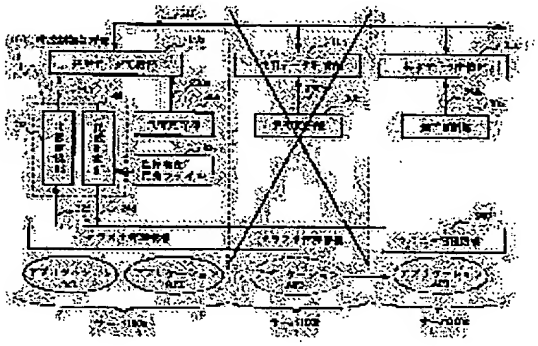
(51)Int.Cl. G06F 15/16
G06F 11/20
G06F 13/00

(21)Application number : 09-350391
(22)Date of filing : 19.12.1997

(71)Applicant : MITSUBISHI ELECTRIC CORP
(72)Inventor : YAMANISHI HIROYUKI
SAITO AKIO

(54) CLUSTER SYSTEM

(57)Abstract:
PROBLEM TO BE SOLVED: To dynamically start/stop an application, decide a start computer, decide a succeeding calculator in accordance with the state change of the computer and the application and to optimize a cluster system in the cluster system.
SOLUTION: A state monitoring part 20 and a load update part 30 take in information showing the state of the computer and the application. Information which is taken in is managed by a common data storage part 10. A constitution control part 40 outputs the instruction of the succession of the application and the change of the succeeding computer to a cluster management mechanism 900 when it receives a trigger from a state monitoring part 20 and a load update part 30.



LEGAL STATUS

[Date of request for examination] 02.02.1998
[Date of sending the examiner's decision of rejection] 22.05.2001
[Kind of final disposal of application other than the examiner's decision of rejection or application converted registration]
[Date of final disposal for application]
[Patent number]
[Date of registration]
[Number of appeal against examiner's decision of rejection]
[Date of requesting appeal against examiner's decision of rejection]
[Date of extinction of right]

(19)日本国特許庁（ＪＰ）　(12)公開特許公報（Ａ）　(11)特許出願公開番号

特開平11－184825

(43)公開日　平成11年(1999)７月９日

(51)Int.Cl.[6]		識別記号	FI		
Ｇ０６Ｆ	15/16	３８０	Ｇ０６Ｆ	15/16	３８０Ｚ
	11/20	３１０		11/20	３１０Ｆ
	13/00	３５５		13/00	３５５

審査請求　有　　請求項の数８　ＯＬ　（全 19 頁）

(21)出願番号　特願平9－350391

(22)出願日　平成９年(1997)12月19日

(71)出願人　000006013
三菱電機株式会社
東京都千代田区丸の内二丁目２番３号

(72)発明者　山西　宏幸
東京都千代田区丸の内二丁目２番３号　三菱電機株式会社内

(72)発明者　斎藤　彰男
東京都千代田区丸の内二丁目２番３号　三菱電機株式会社内

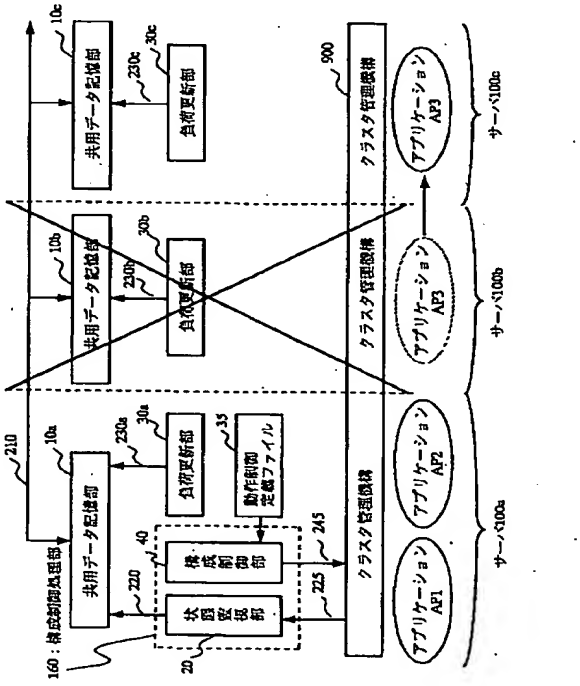
(74)代理人　弁理士　宮田　金雄　（外２名）

(54)【発明の名称】　クラスタシステム

(57)【要約】

【課題】　クラスタシステムにおいて、計算機及びアプリケーションの状態変化に応じてアプリケーションの起動、停止、起動計算機の決定及び引き継ぎ計算機の決定を動的に行い、クラスタシステムの最適化を図る。

【解決手段】　状態監視部２０、負荷更新部３０が計算機及びアプリケーションの状態を示す情報を取り込む。取り込んだ情報は、共用データ記憶部１０で管理する。構成制御部４０は、トリガを状態監視部２０、負荷更新部３０から受け取ると、アプリケーションの引き継ぎや引き継ぎ計算機の変更などの指示をクラスタ管理機構９００に出す。



1

【特許請求の範囲】

【請求項1】　１つ以上のアプリケーションを実行可能な複数台のサーバをメンバーとして構成されるクラスタシステムであって、上記複数台のサーバのいずれか1台に上記アプリケーションを実行させるとともに、上記1台のサーバが上記アプリケーションを実行不可能である時に他のサーバに上記アプリケーションの実行を引き継がせるクラスタシステムにおいて、

以下の要素を有するクラスタシステム

（ａ）上記複数台のサーバから共通して利用可能な共用データとして、上記複数台のサーバの状態を示すサーバ状態情報と上記複数台のサーバの負荷を示す負荷情報と上記アプリケーションの動作に関するアプリケーション動作情報とを記憶する共用データ記憶部、（ｂ）上記複数台のサーバの状態を取得して上記サーバ状態情報を更新する状態監視部、（ｃ）上記複数台のサーバの負荷を取得して上記負荷情報を更新する負荷更新部、（ｄ）上記共用データ記憶部を参照して、上記アプリケーションと上記複数台のサーバとの対応を動的に制御する構成制御部。

【請求項2】　上記共用データ記憶部は、上記クラスタシステムの制御に関する規約を定義する規約情報を記憶するとともに、上記構成制御部は、上記クラスタシステムの起動の際のクラスタ構成時に上記規約情報と上記サーバ状態情報とを参照して上記アプリケーション動作情報を作成することを特徴とする請求項1に記載のクラスタシステム。

【請求項3】　上記構成制御部は、上記クラスタシステムのクラスタ構成後に新たなサーバが起動してメンバーとして上記クラスタシステムに参加した時、上記サーバ状態情報と上記アプリケーション動作情報とを参照して上記アプリケーションの引継ぎを行うことを特徴とする請求項1に記載のクラスタシステム。

【請求項4】　上記構成制御部は、上記クラスタシステムのクラスタ構成後にメンバーの移動が発生した時、上記サーバ状態情報と上記アプリケーション動作情報とを参照して上記アプリケーションを引き継ぐ引継ぎサーバを決定して上記アプリケーション動作情報を更新することを特徴とする請求項1に記載のクラスタシステム。

【請求項5】　上記構成制御部は、上記複数のサーバのいずれかの負荷状態に変化が起こった時、上記サーバ負荷情報と上記アプリケーション動作情報とを参照して上記アプリケーションの引継ぎを行うことを特徴とする請求項1に記載のクラスタシステム。

【請求項6】　上記規約情報は、ユーザが設定する定義ファイルをもとに作成されることを特徴とする請求項2に記載のクラスタシステム。

【請求項7】　上記クラスタシステムは、パターン化されたテンプレートを予め用意し、用意したテンプレートのいずれかをもとに上記規約情報を作成することを特徴

2

とする請求項2に記載のクラスタシステム。

【請求項8】　上記負荷更新部は、所定のタイミングで定期的に上記複数台のサーバの負荷を取得して上記負荷情報を更新し、上記構成制御部は、上記負荷情報を参照してアプリケーション引継ぎに際し、上記アプリケーションを引き継ぐ引継ぎサーバを決定して上記アプリケーションの引継ぎを行うことを特徴とする請求項1に記載のクラスタシステム。

【発明の詳細な説明】

【０００１】

【発明の属する技術分野】この発明は、分散した複数の計算機を一まとまりにして構成したクラスタシステムに関するものである。特に、クラスタシステムでのアプリケーションの引き継ぎに関するものである。

【０００２】

【従来の技術】クラスタシステムにおいては、複数の計算機で1つ以上のアプリケーションを実行し、アプリケーション、もしくはアプリケーションを実行中の計算機が異常停止した場合、クラスタ内の他の計算機でアプリケーションを引き継いで実行する。この、アプリケーションを引き継いで実行する計算機を引継ぎ計算機と呼ぶ。引継ぎ計算機は、クラスタシステム内で、予め設定される。このような、アプリケーションの切り替えを示した従来の技術として、特開平6－２８２０６号公報に開示された「クラスタシステムのデータ処理ステーション障害時の復旧方式」がある。この技術は、運用側の故障により無条件に予備側に切り替えるものである。だが、その後の新しい運用側の障害には対処できないという欠点があった。

【０００３】

【発明が解決しようとする課題】このように、従来のクラスタの構成制御方式では、自動引継ぎ制御しか行われておらず、異常停止した計算機が復帰した等の事象によってクラスタ構成計算機メンバーが変更された場合、アプリケーションの引継ぎ計算機の再設定などは、システム管理者がコマンド等を実行して行う必要があった。また、アプリケーションが計算機に引き継がれる際、計算機やアプリケーションの状況に対応した引継ぎ計算機の自動変更処理が行われておらず、重複障害発生時に1つの計算機にアプリケーションが集中してしまうこともある。

【０００４】この発明は、上記のような問題点を解決するためになされたものであり、計算機及びアプリケーションの状態変化に応じて、アプリケーションの起動、停止、起動計算機の決定を自動的に行うクラスタシステムを得ることを目的としている。また、計算機及びアプリケーションの状態変化に応じて、動的に引継ぎ計算機の決定を行い、クラスタ構成を最適に制御するクラスタシステムを得ることを目的としている。また、予め形式の決まった運転形態を簡単に設定することができる機能を

備えたクラスタシステムを得ることを目的としている。

【０００５】

【課題を解決するための手段】この発明に係るクラスタシステムは、１つ以上のアプリケーションを実行可能な複数台のサーバをメンバーとして構成されるクラスタシステムであって、上記複数台のサーバのいずれか1台に上記アプリケーションを実行させるとともに、上記1台のサーバが上記アプリケーションを実行不可能である時に他のサーバに上記アプリケーションの実行を引き継がせるクラスタシステムにおいて、以下の要素を有することを特徴とする。

（ａ）上記複数台のサーバから共通して利用可能な共用データとして、上記複数台のサーバの状態を示すサーバ状態情報と上記複数台のサーバの負荷を示す負荷情報と上記アプリケーションの動作に関するアプリケーション動作情報とを記憶する共用データ記憶部、（ｂ）上記複数台のサーバの状態を取得して上記サーバ状態情報を更新する状態監視部、（ｃ）上記複数台のサーバの負荷を取得して上記負荷情報を更新する負荷更新部、（ｄ）上記共用データ記憶部を参照して、上記アプリケーションと上記複数台のサーバとの対応を動的に制御する構成制御部。

【０００６】上記共用データ記憶部は、上記クラスタシステムの制御に関する規約を定義する規約情報を記憶するとともに、上記構成制御部は、上記クラスタシステムの起動の際のクラスタ構成時に上記規約情報と上記サーバ状態情報とを参照して上記アプリケーション動作情報を作成することを特徴とする。

【０００７】上記構成制御部は、上記クラスタシステムのクラスタ構成後に新たなサーバが起動してメンバーとして上記クラスタシステムに参加した時、上記サーバ状態情報と上記アプリケーション動作情報とを参照して上記アプリケーションの引継ぎを行うことを特徴とする。

【０００８】上記構成制御部は、上記クラスタシステムのクラスタ構成後にメンバーの移動が発生した時、上記サーバ状態情報と上記アプリケーション動作情報とを参照して上記アプリケーションを引き継ぐ引継ぎサーバを決定して上記アプリケーション動作情報を更新することを特徴とする。

【０００９】上記構成制御部は、上記複数のサーバのいずれかの負荷状態に変化が起こった時、上記サーバ負荷情報と上記アプリケーション動作情報とを参照して上記アプリケーションの引継ぎを行うことを特徴とする。

【００１０】上記規約情報は、ユーザが設定する定義ファイルをもとに作成されることを特徴とする。

【００１１】上記クラスタシステムは、パターン化されたテンプレートを予め用意し、用意したテンプレートのいずれかをもとに上記規約情報を作成することを特徴とする。

【００１２】上記負荷更新部は、所定のタイミングで定期的に上記複数台のサーバの負荷を取得して上記負荷情報を更新し、上記構成制御部は、上記負荷情報を参照してアプリケーション引継ぎに際し、上記アプリケーションを引き継ぐ引継ぎサーバを決定して上記アプリケーションの引継ぎを行うことを特徴とする。

【００１３】

【発明の実施の形態】実施の形態1．最初に、この発明のクラスタシステムが実現する機能の概要を説明する。

１．実現機能の概要

（ａ）計算機の初期起動時、即ち、個々の計算機が電源投入され、クラスタの初期構成が行われる時、クラスタ内のアプリケーションの引継ぎ計算機を自動制御し、決定する機能。

（ｂ）計算機の起動時、アプリケーションを自動的に引き継がせる機能。

（ｃ）計算機異常停止時又は異常停止からの復帰時など、クラスタを構成する計算機に変更があった際、クラスタ内のアプリケーションの引継ぎ計算機を自動制御して変更する機能。

（ｄ）計算機が通常、起動している状態（ＣＰＵ、メモリ、ネットワーク）の負荷を監視し、その状態の変化に応じて、アプリケーションを自動制御する機能。

（ｅ）プログラミングを行うことなく、定義ファイルを設定するだけで、構成の再構築制御を行う機能。

（ｆ）ホットスタンバイ、ロードシェア、Ｎ：１バックアップ型など、予め形式の決まった運転形態を簡単に設定することができる機能。

（ｇ）アプリケーションを自動で引き継がせる場合、各計算機の負荷も考慮し、引継ぎ計算機を決定する機能。

【００１４】次に、図を用いてこの発明のクラスタシステムの構成を説明する。図1は、この発明のクラスタシステムの構成の一例を示すブロック図である。図において、９００は既存のクラスタ管理機構である。この発明のクラスタシステムでは、サーバ１００ａ～１００ｃ（計算機ともいう）の状態及びアプリケーションの状態に関する情報をサーバ１００ａの状態監視部２０でクラスタ管理機構９００から取り込み、それらの情報を共用データ記憶部１０ａに記憶する（矢印２２０）。この共用データ記憶部１０ａは、矢印２１０に示すように、他のサーバの共用データ記憶部１０ｂ、共用データ記憶部１０ｃにコピーして同じ状態に置くよう管理されるものとする。また、負荷更新部３０ａ～３０ｃは、サーバ１００ａ～１００ｃの負荷に関する情報を得て共用データ記憶部１０ａに記憶する（２３０ａ～２３０ｃ）。構成制御部４０は、システムの状態変化を状態監視部２０から受けると、又はシステムの負荷の変化を各計算機に存在する負荷更新部３０ａ～３０ｃから受けると、アプリケーションの引継ぎや、引継ぎ計算機の変更などの指示をクラスタ管理機構９００へ出す。状態監視部２０と構成制御部４０を構成制御処理部１６０と呼ぶ。ここで

5

は、構成制御処理部１６０がサーバ１００ａ上で動作する場合を示しているが、サーバ１００ｂやサーバ１００ｃ上で動作しても構わない。

【００１５】以下に、前述したクラスタシステムの機能を実現する実現手段について説明する。

２．実現手段の概要

（ａ）クラスタの初期起動の通知を受けた構成制御部40は、実行中のアプリケーションの引継ぎ計算機を、アプリケーションを実行していない計算機に設定する。

（ｂ）ある計算機が動作中のクラスタに参加してきた時、その通知を受けた構成制御部４０は、その計算機上で動作すべきアプリケーションを検索し、動作すべきアプリケーションが検索されれば、そのアプリケーションを引き継がせる指示を出す。

（ｃ）クラスタに参加中の計算機が停止したり、動作中のクラスタにある計算機が参加した場合、その通知を受けた構成制御部４０は、引継ぎ計算機を変更する指示を出す。

（ｄ）各計算機の負荷状態（ＣＰＵ、メモリ、ネットワーク）を監視し、その状態を管理する機能を構成制御部４０に持たせる。実行中のアプリケーションをその負荷の状況に応じて、引継ぎ計算機に引き継がせる指示を出す。

（ｅ）計算機が異常停止した時、停止していた計算機が起動した時のアプリケーションの動作について定義した制御動作定義ファイル３５を管理コマンドを用いて、各計算機へ配布／配置する。構成制御部４０は、各要素に変化があった場合、その情報に従った自動制御を行う。

（ｆ）予めパターン化されたクラスタの自動構成制御を行うための定義情報のテンプレートを用意し、ユーザに開放する。テンプレートをもとに、制御動作定義ファイル３５が作成される。

（ｇ）負荷更新部３０ａ～３０ｃが定期的に負荷の情報を取り込み、取り込んだ情報を共用データ記憶部１０ａ～１０ｃへと反映する。各計算機の負荷情報を管理する管理機能を持った構成制御部４０が、その負荷情報が示す負荷状況に応じて最適なアプリケーションの引継ぎを指示する。図2に、アプリケーションの引継ぎの例を示す。図2は、サーバ１００ｂが異常停止し、アプリケーションＡＰ３がサーバ１００ｃに引き継がれた状態を示している。

【００１６】３．ハードウェア構成

図3は、この発明のクラスタシステムのハードウェア構成の一例を示す図である。ハードウェアは、各ローカルディスク２００ａ～２００ｎを持つN台のサーバ１００ａ～１００ｎ、各サーバ間の通信路であるネットワーク１０００から構成される。

【００１７】４．ソフトウェア構成

図4，図5は、マルチサーバ上のソフトウェアの位置付けを説明する図である。ソフトウェアは、大別して下記

6

部位から構成される。

（ａ）既存のクラスタ管理機構９００

アプリケーションパッケージを、定義された情報に従ってサーバに割り付け、実行させるクラスタ管理機構である。クラスタ管理機構９００は、アプリケーション実行中のサーバが異常停止した場合、アプリケーション毎に定義されている別サーバで再起動するなどの制御を行う。本実施の形態では、ヒューレットパッカード社（米国のＨｅｗｌｅｔｔ－Ｐａｃｋａｒｄ　Ｃｏｍｐａｎｙ、ＨＰ社ともいう。Ｈｅｗｌｅｔｔ－Ｐａｃｋａｒｄ　Ｃｏｍｐａｎｙ、ＨＰはＨＰ社の商標又は登録商標）のＭＣ／ＳｅｒｖｉｃｅＧｕａｒｄをクラスタ管理機構として使用する場合を例にとって説明する。クラスタ管理機構は、既存の製品であり、本発明以前の従来からある技術である。

【００１８】（ｂ）構成制御処理部１６０

クラスタを構成する全てのサーバ上で動作可能な構成制御処理部１６０がある。構成制御処理部１６０は、状態監視部２０と構成制御部４０からなる。構成制御処理部１６０は、クラスタを構成する全てのサーバのうち、いずれかのサーバ上で動作する。構成制御処理部１６０は、全てのサーバで動作可能なもので、クラスタ管理機構９００に対して予め設定することにより、下記のようにクラスタ管理機構９００によって起動される。

・初期起動

クラスタ管理機構９００の設定（起動サーバの優先順位指定）に従い、現状構成内の高優先順位のサーバで起動される。

・動作サーバ停止時の再起動

動作中のサーバが異常停止した場合、クラスタ管理機構の設定に記述された次優先順位のサーバで再起動され、引き続き動作する。図5は、サーバ１００ａが停止し、構成制御処理部１６０がサーバ１００ｂで引き続き動作する場合を示している（矢印２１９）。この初期／再起動のメカニズムは、既存のクラスタ管理機構が行う。構成制御処理部１６０は、前述した機能を実現し、アプリケーションの移動先サーバの決定などを行う。決定された移動先のサーバの指示は、クラスタ管理機構９００に対して行われる。また、アプリケーションの起動、移動等の制御は、クラスタ管理機構９００により行われる。クラスタ管理機構９００は、構成制御処理部１６０からの指示に従い動作する。即ち、この実施の形態の構成制御処理部１６０は、クラスタ管理機構９００の存在を前提としている。構成制御処理部１６０は、直接アプリケーションを起動したり、アプリケーションの引き継ぎを行ったりすることはない。

【００１９】（ｃ）共用データ記憶部１０ａ～１０ｃ

ネットワークを介した共用メモリ機能（あるサーバで書き込んだ内容を別サーバに反映することで、別サーバがその内容を参照できる機能）をサポートする記憶部であ

る。構成制御処理部１６０が別サーバで再起動された場合の引き継ぎデータとして使用される。この共用メモリ機能の実現方式は問わないものとする。この実施の形態では、この共用メモリ機能にサポートされた共用データ記憶部を用いる場合について説明する。データの共用方式には、特殊な装置や共用ディスクを使用した様々な方式が考えられるが、本実施の形態は、ネットワークを介した方式で記述する。

【００２０】（ｄ）アプリケーションプロセス（アプリケーションＡＰ１～ＡＰ３）

構成制御処理部１６０の動作結果として、障害時のバックアップサーバ設定などが行われるユーザのアプリケーションプログラム。実際のサーバ起動は、クラスタ管理機構が行う。本発明のクラスタシステムは、クラスタ管理機構９００に対して指示を出す。図5の矢印２３４は、アプリケーションＡＰ１、アプリケーションＡＰ２がサーバ１００ａの停止に伴い、サーバ１００ｃで引き継ぎ実行される状態を示している。その際、構成制御処理部１６０は、共用データ記憶部１０ｂに定義されている情報を参照してクラスタ管理機構９００に対してアプリケーションの起動指示を出し（２４６）、指示を受け取ったクラスタ管理機構９００は、２３２に示すように、アプリケーションＡＰ１，ＡＰ２をそれぞれ起動する。構成制御処理部１６０は、２２６に示すように、クラスタ管理機構９００から各サーバ及びアプリケーションの状態を取得する。

【００２１】５．システム動作概要

次に、クラスタシステムの動作について説明する。図6は、クラスタシステムの動作概要図である。大きな流れは、以下のようになる。

（１）常時実行の状況収集

常時、下記情報を取り込み、共用データ記憶部に書き込み、他サーバの共用データ記憶部に反映を行う。

（ａ）サーバ状態

状態監視部２０は、クラスタにサーバが組み込まれているか否かの状態をクラスタ管理機構９００より取り込み（矢印２２５）、全サーバの状態を管理する。サーバの突然の異常停止や異常停止後の立ち上がりなどを検知する。共用データ記憶部は、２１０に示すように、他サーバに反映を行う。

（ｂ）各サーバの負荷状態

各サーバの負荷更新部３０ａ～３０ｃは、自サーバの負荷をそれぞれ取り込み、共用データ記憶部へ書き込む。書き込まれた各サーバの負荷は、それぞれ他サーバへ反映される。他サーバへお互いに反映することにより、各サーバの共用データ記憶部は、全てのサーバの負荷状態を記憶する。

【００２２】（２）アプリケーションのサーバ割り付けの管理、制御

構成制御処理部の総合的な管理として、構成制御部が下記に示す各アプリケーションプログラムの実行状態の管理、制御を行う。

１）管理

各サーバの状態と合わせて、各アプリケーションの下記状態を常に管理している。

（ａ）各アプリケーションの動作サーバ

アプリケーションの動作サーバとは、そのアプリケーションが動作しているサーバである。

（ｂ）各アプリケーションのバックアップサーバ

各アプリケーションが動作しているサーバが停止した場合、どのサーバで再起動させるかを定義する情報である。再起動させるサーバをバックアップサーバという。

２）制御

後述されるタイミング（構成サーバ変更、定周期）で、各アプリケーションの動作サーバとバックアップサーバを再定義し、クラスタ管理機構に動作サーバとバックアップサーバの再定義の指示を行う。なお、動作サーバ、バックアップサーバの再定義は、予め定義された制御動作定義ファイルに沿って行われる。制御動作定義ファイル３５には、下記の２つの記述方式がある。

（ａ）規約記述型

サーバ状態に応じた割り付け指示などの規約の詳細をシステム管理者が予め記述する。状況に応じた複雑な指示が可能という長所がある。だが、システム管理者の負荷が重いという欠点もある。

（ｂ）モデル指定型

クラスタシステムが用意する代表的な形態のモデルをシステム管理者が指定する簡易な設定方式である。設定が簡単に短時間で行え、システム管理者の負荷を大幅に軽減するという長所がある。例えば、予め用意されたモデルから二重系ホットスタンバイ、多重系N対1バックアップ型などのモデルを指定できる。

【００２３】（３）初期立ち上げ時の動作

初期立ち上げ時には、構成制御部４０が制御動作定義ファイルをもとに作成された共用データ記憶部の制御動作規約情報テーブル（後述）を参照して動作サーバとバックアップサーバを定義し、各アプリケーションの起動要求をクラスタ管理機構９００に伝える。

【００２４】（４）構成サーバ変更時の動作

サーバの異常停止及び異常停止後の再立ち上げにより、構成サーバが変更した時、状態監視部２０からのトリガ２５２で、各アプリケーションのバックアップサーバの再定義を構成制御部４０が行う。構成制御部４０は、再定義したバックアップサーバをクラスタ管理機構９００に伝える。なお、バックアップサーバの再設定先は、サーバ構成及び各サーバの負荷状況をもとに決定される。

【００２５】（５）定周期による動作

負荷更新部３０ａからのトリガ２５４で、構成制御部４０は、定期的に各サーバの負荷をチェックし、チェックの結果により、各アプリケーションのバックアップサー

バを低負荷のサーバに再設定し、アプリケーション移動時に特定サーバに負荷が集中することを防ぐ。

【００２６】（６）サーバ異常時のアプリケーションの移動動作

実際に、サーバの異常停止が発生した場合、構成制御処理部１６０により設定されたバックアップサーバでアプリケーションを再起動させるために、クラスタ管理機構９００にアプリケーションの再起動の指示を出す。指示を受け取ったクラスタ管理機構9００がアプリケーションを再起動する。以上のように、この発明のクラスタシステムは動作する。

【００２７】６．サーバ状態の監視

次に、状態監視部２０が行うサーバ状態の監視について説明する。図7は、サーバ状態の監視を説明する図である。状態監視部２０によるサーバ状態の取得は、このクラスタシステムが使用するクラスタ管理機構９００の提供する状態取得機能を使用して行う。状態監視部２０は、クラスタ管理機構９００からポーリング形式でサーバ状態を取得する。本実施の形態では、今回取得したサーバ状態の情報と前回取得したサーバ状態の情報とを比較し、状態変化を検知する方式を採用している。主たる動作は、以下の通りである。

（ａ）クラスタ管理機構９００の提供するサーバ状態取得コマンドを使用し、ポーリング形式で全サーバの状態を取得する（矢印２６０）。

（ｂ）初回の取得時、共用データ記憶部のサーバ状態情報テーブル１２ａに記憶し、他サーバのサーバ状態情報テーブル１２ｂ、１２ｃへ反映を行う（矢印２１０）。

（ｃ）２回目以降の取得時（前回の取得結果を記録済みの時）は、記録情報と比較する。状態変化を検知すると、共用データ記憶部のサーバ状態情報テーブル１２ａに取得した状態及び変更マーク（切り離し／組み込みマーク）を記憶し、他サーバのサーバ状態情報テーブル１２ｂ、１２ｃへ反映を行う（矢印２１０）。

（ｄ）初回及び変更時、構成制御部へ制御トリガを指示する。トリガは、ＯＳ（Ｏｐｅｒａｔｉｎｇ　Ｓｙｓｔｅｍ）が提供するメッセージ送信機能（プログラム間通信機能）を使用する。

【００２８】７．負荷状況の監視

次に、負荷更新部３０ａ～３０ｃが行う負荷状況の監視について説明する。図8は、負荷状況の監視を説明する図である。各サーバの負荷更新部が自サーバの負荷を取り込み、共用データ記憶部１０ａ～１０ｃに書き込み、他サーバへ反映を行う。動作は、以下の通りである。

（ａ）負荷更新部３０ａ～３０ｃは、定周期で起動される。

（ｂ）負荷更新部３０ａ～３０ｃは、ＯＳが提供するＣＰＵ負荷情報取り込みコマンドで、ＯＳツール群９１０ａ～９１０ｃより自サーバのＣＰＵ負荷を取り込み（２６２ａ～２６２ｃ）、共用データ記憶部のサーバ負荷情報テーブル１３ａ～１３ｃに書き込み（２６８ａ～２６８ｃ）、他サーバの共用データ記憶部１０ａ～１０ｃへそれぞれ反映する。

（ｃ）構成制御部４０が動作するサーバ上で動作する負荷更新部の場合、構成制御部４０へバックアップサーバチェックのトリガを送る（矢印２６４）。

【００２９】８．アプリケーション状況の管理、制御

次に、構成制御部４０が行うアプリケーション状況の管理、制御について説明する。構成制御部４０は、各アプリケーションの状態を管理し、状態監視部２０又は負荷更新部３０からのトリガによって、各アプリケーションのバックアップサーバの変更などを行う。

【００３０】（１）管理データ

管理、制御するための必要な管理データは、以下のものである。なお、各データはすべて共用データ記憶部１０ａ～１０ｃに確保され、更新時は全てのサーバに反映される。従って、各サーバは同一イメージの管理データを有することになる。

【００３１】（ａ）制御動作規約情報テーブル１１

図9に、制御動作規約情報テーブル１１を示す。制御動作規約情報テーブル１１は、サーバの構成に応じ、各アプリケーションの動作サーバとバックアップサーバの割り付け定義が記述された動作規約定義である。定義自体は、定義ファイル制御動作定義ファイル３５に記述され、定義ファイル制御動作定義ファイル３５の内容が共用データ記憶部に制御動作規約情報テーブル１１ａ～１１ｃとしてクラスタの初期構成時に展開される。制御動作規約情報テーブル１１は、クラスタシステム動作中に書き替えられることはなく、固定値である。各規約の意味は、以下の通りである。

・動作サーバ１１０５

クラスタの初期立ち上げ時にアプリケーションを動作させるサーバ名である。

・バックアップサーバ１１０７

初期立ち上げ時、設定されるバックアップサーバ名である。未定義時、バックアップサーバを割り付けない。

・固定指示１１０９

定期負荷によりバックアップサーバの切替を行うか否かの指定である。切り替えを行う時“ＯＦＦ”、切り替えを行わない時“ＯＮ”とする。

・自動移動モード１１１１

同一サーバで複数のアプリケーションが動作している状態で、他のサーバが復旧した時、その復旧したサーバに複数のアプリケーションのうちのいずれかのアプリケーションを移動させるか否かの指定である。否の時、復旧したサーバへのアプリケーションの移動は行わずに、復旧したサーバへバックアップサーバを割り付ける処理となる。固定指示１１０９、自動移動モード１１１１の2つの規約は、アプリケーションを移動させるためには、そのアプリケーションを終了させなければならないの

で、業務の内容により途中で停止させることが望ましくない場合への対応として用意されている。

【００３２】（ｂ）サーバ状態情報テーブル１２

図１０に、サーバ状態情報テーブル１２を示す。各サーバの正常／異常などの状態を管理するためのもので、状態監視部２０が共用データ記憶部に作成、更新する。各パラメータの意味は、下記の通りである。

・サーバ状態１２０３

サーバがクラスタに組み込まれているか否か（正常／切り離し）の状態を示す。

・検出状態１２０５

状態変化のあったサーバについて、検出した状態の変化が“クラスタからの切り離し”か“クラスタへの組み込み”かを示す。

・処理モード１２０７

サーバ状態の変化に伴う変更処理が処理中か否かを示す。構成制御処理部１６０の構成制御動作途中にサーバの移動が発生した場合、サーバ移動後のリカバリに使用する。

【００３３】（ｃ）サーバ負荷情報テーブル１３

図１１に、サーバ負荷情報テーブル１３を示す。各サーバのＣＰＵ負荷情報を管理するためのもので、負荷更新部が定期的に更新する。サーバ負荷情報テーブル１３は、共用データ記憶部に確保される。各パラメータの意味は、下記の通りである。

・負荷情報１３０３

各サーバのＣＰＵ負荷（％）を示す。

【００３４】（ｄ）アプリケーション動作情報テーブル１４

図１２に、アプリケーション（ＡＰ）動作情報テーブル１４を示す。各アプリケーションの動作サーバとバックアップサーバを管理するもので、構成制御部４０が作成、更新する。また、構成制御部４０は、アプリケーション動作情報テーブル１４の更新内容に従って、クラスタ管理機構９００に実行指示を行う。各パラメータの意味は、下記の通りである。

・動作サーバ名１４０３

アプリケーションが実際に動作しているサーバ名が保持される。

・バックアップサーバ名１４０５

実際にバックアップサーバとして割り付けられているサーバ名が保持される。

・処理中フラグ１４０７

アプリケーションに対する動作サーバやバックアップサーバの変更処理が、処理中か否かを示す。構成制御処理部１６０の構成制御動作途中に、サーバの移動が発生した場合、サーバ移動後のリカバリに使用する。

【００３５】（２）構成制御部４０の動作トリガ

構成制御部４０の動作トリガには、下記のものがある。

・初期立ち上げ

システムが開始されたタイミングで、状態監視部２０から通知される。

・構成変更時

サーバ異常停止、再立ち上げ、操作指示によるサーバ切り離し／組み込みにより、サーバの状態が変更されたタイミングで、状態監視部２０から通知される。

・定期

負荷更新部が定期的に行う自サーバ負荷更新タイミングで、構成制御部４０が動作しているサーバの負荷更新部３０から通知を受けとる。構成制御部４０は、動作トリガの通知をＯＳ提供のメッセージ通知機構（プログラム間通信機能）で受けとる。構成制御部４０は、メッセージの通知を常に待ち、受けとると要因に沿った処理を行う。

【００３６】（３）初期立ち上げ時の構成制御部４０の動作

図１３は、構成制御部４０の初期立ち上げ時の動作を示す流れ図である。構成制御部４０は、状態監視部２０から何らかのメッセージの送付を受けると、Ｓ１０１において、メッセージを取り込み、メッセージの要因が“初期立ち上げ”かどうか判定する。ＮＯの場合、Ｓ１０３において、他の要求のメッセージかどうかチェックを行う。ＹＥＳの場合、Ｓ１０５において、サーバ状態情報テーブルから現状のサーバ構成を取得し、本構成に対応する制御動作規約情報テーブル１１を参照し、各アプリケーションの動作サーバとバックアップサーバを決定し、アプリケーション動作情報テーブルを作成する。次に、Ｓ１０７で、クラスタ管理機構９００に、各アプリケーションをアプリケーション動作情報テーブルに定義された動作サーバで起動するよう指示を行う。次に、Ｓ１０９において、クラスタ管理機構９００にアプリケーションのバックアップサーバの設定を行うよう指示を出す。Ｓ１１１において、全アプリケーションの処理が終了したか否かを判定し、終了するまでＳ１０５～Ｓ１０９の処理を繰り返す。このように、全アプリケーションに対し、上記起動／設定処理を行う。Ｓ１１１の判定で全アプリケーション処理が終了すると、Ｓ１１３でメッセージ待ちに戻る。この状態で、アプリケーション動作サーバが異常停止すると、クラスタ管理機構９００は、構成制御部４０の指示により設定されているバックアップサーバでアプリケーションを再起動する。

【００３７】（４）構成変更時の構成制御部４０の動作

図１４から図１８は、サーバ構成が変更された時の構成制御部４０の動作を示す流れ図である。

１）動作サーバ切り離し時

サーバの異常停止やユーザ指示によりサーバがクラスタから削除された場合、構成制御部４０は、図１４，図１５に示した流れ図に従い、下記動作を行う。構成制御部４０は、状態監視部２０からメッセージの送付を受けると、Ｓ１２１において、送付されたメッセージを取り込

み、取り込んだメッセージが“切り離し要求”かどうか判定する。判定した結果、“切り離し要求”でなければ、Ｓ１２３において、他の要因のチェックを行う。“切り離し要求”であった場合には、Ｓ１２５において、サーバ状態情報テーブル１２から切り離し要求のあるサーバ（検出状態１２０５が切り離しのサーバ）をサーチする。検出したら、サーバ状態情報テーブル１２の処理モード１２０７を処理中にセットする（Ｓ１２９）。検出しない場合は、元のメッセージ待ちに戻る（Ｓ１２７）。

【００３８】次に、Ｓ１３１において、アプリケーション動作情報テーブルから該当サーバで動作していたアプリケーションをスキャンする。次に、Ｓ１３３において、検出したアプリケーションのアプリケーション動作情報テーブルの処理中フラグ１４０７をセットし、実際に動作しているサーバ名をクラスタ管理機構９００から取り込み、取り込んだサーバ名をアプリケーション動作情報テーブル１４の動作サーバ名１４０３にセットする。この時点では、切り離し要求のあるサーバ上で動作していたアプリケーションは、既にクラスタ管理機構９００によりアプリケーション動作情報テーブル１４で定義されているバックアップサーバで再起動しているはずであるが、構成制御部４０は、クラスタ管理機構９００から実際の動作サーバ名を取り込む。

【００３９】実際に動作しているサーバが存在しない（アプリケーション停止＝バックアップ未定義又は重複障害）場合は、図１５のＳ１３９に示すように、バックアップサーバ無しを選択する。

【００４０】アプリケーションが動作中の場合、制御動作規約情報テーブルの同一サーバ構成の項を参照し（図１４、Ｓ１３５）、アプリケーションのバックアップサーバが未定義の場合、図１５のＳ１３９に示すように、バックアップサーバ無しを選択する。アプリケーションのバックアップサーバの未定義は、故障により動作サーバ数が特定台数に満たない場合、サーバの処理能力上の問題から、重要でないアプリケーションを切り捨て、重要なアプリケーションの動作を保証するため等に使用される。

【００４１】バックアップサーバが定義されている場合、アプリケーション動作情報テーブル／サーバ状態情報テーブルから、アプリケーションが動作していない正常サーバをサーチする（図１５、Ｓ１３７）。検出した場合、検出した正常サーバをバックアップサーバに決定する（図１５、Ｓ１４３）。

【００４２】検出しない場合、サーバ負荷テーブルから最も低負荷なサーバをバックアップサーバに選択する（図１５、Ｓ１４１）。

【００４３】次に、Ｓ１４５で、アプリケーション動作情報テーブルのバックアップサーバ名を選択したバックアップサーバに変更し、他サーバに反映を行う。その

後、Ｓ１４７で、クラスタ管理機構９００にバックアップサーバの変更指示を行う。完了後、アプリケーション動作情報テーブル１４の処理中フラグ１４０７をリセットし、他サーバへ反映を行う（Ｓ１４９）。Ｓ１３１～Ｓ１４９の処理を該当する全アプリケーションに対して繰り返し行う。

【００４４】全アプリケーションに対する処理が完了したら、サーバ状態情報テーブル１２の該当サーバのサーバ状態１２０３を切り離し状態にし、検出状態１２０５をクリア、処理モード１２０７をリセットし、他サーバへ反映する（Ｓ１５１）。

【００４５】２）サーバ再組み込み時

異常サーバの再立ち上げやコマンド指示によるサーバ組み込みで、クラスタにサーバが再組み込みされた場合、構成制御部４０は、図１６から図１８に示す流れ図に従い、下記動作を行う。

【００４６】構成制御部４０は、状態監視部２０からメッセージを送付されると、Ｓ１６１において、送付されたメッセージを取り込み、メッセージが“組み込み要求”かどうか判定する。“組み込み要求”でなかった時には、Ｓ１６３において、他の要求のチェックを行う。“組み込み要求”であった場合には、Ｓ１６５において、サーバ状態情報テーブル１２から組み込み要求のあるサーバ（検出状態１２０５が組み込み）をサーチする。検出したら、サーバ状態情報テーブル１２の処理モード１２０７を処理中にセットする（Ｓ１６９）。検出しない場合は、元のメッセージ待ちに戻る（Ｓ１６７）。

【００４７】次に、アプリケーション動作情報テーブル１４をサーチし、動作していないアプリケーション（動作サーバ名１４０３に登録のないアプリケーション）をサーチする（Ｓ１７１）。検出した場合、下記を行う。アプリケーション動作情報テーブル１４の処理中フラグ１４０７を処理中にし、動作サーバ名１４０３を組み込みサーバに設定し、他サーバへ反映する（Ｓ１７３）。次に、Ｓ１７５で、アプリケーションの起動指示をクラスタ管理機構９００に対して行う。完了後、アプリケーション動作情報テーブル１４の処理中フラグ１４０７をリセットし、他サーバへ反映する。該当するアプリケーションに対し処理を行ったら、元のメッセージ待ちへ戻る（Ｓ１７９）。

【００４８】Ｓ１７１の判定で、動作サーバなしのアプリケーションが無い場合は、図１７のＳ１８１において、同一サーバで重複動作しているアプリケーションをアプリケーション動作情報テーブル１４からサーチする。存在する場合、下記のアプリケーション移動処理を行う。制御動作規約情報テーブル１１を参照し、同一サーバで重複動作しているアプリケーションの内、同一サーバ構成時、自動移動モードがＯＮのものをサーチする（Ｓ１８３）。存在しない場合は、アプリケーション移

動処理はスキップし、図18のS193に進む。検出した場合、図17のS185において、そのアプリケーションのアプリケーション動作情報テーブル14の処理中フラグ1407をセットし、動作サーバ名1403を組み込みサーバに変更し、他サーバに反映する。その後、クラスタ管理機構900に、該当アプリケーションの移動(停止、起動サーバ設定、起動)指示を行う(S187)。完了後、アプリケーション動作情報テーブル14の処理中フラグ1407をリセットし、他サーバに反映する(S189)。処理終了後、元のメッセージ待ちへ戻る(S191)。

【0049】上記のアプリケーション移動処理が行われなかった場合、図18のS193において、各アプリケーションのバックアップサーバ名1405を組み込みサーバに変更する。本処理は、前述同様、アプリケーション動作情報テーブル14の処理中フラグ1407を処理中にし、バックアップサーバ名1405を書き換え、他サーバに反映する。次に、S195において、クラスタ管理機構900に、バックアップサーバ名変更指示を行う。次に、S197において、アプリケーション動作情報テーブル14の処理中フラグ1407をリセットし、他サーバに反映する。

【0050】最後に、サーバ状態情報テーブル12のサーバ状態1203を正常に書き換え、検出状態1205をクリア、処理モード1207をリセットし、他サーバへ反映する(S199)。

【0051】(5)定期負荷更新時の動作

負荷更新部からのトリガがあった場合、構成制御部40は、図19に示す流れ図に従い、下記動作を行う。構成制御部40は、負荷更新部30からメッセージを送付されると、S211で送付されたメッセージを取り込み、"定期負荷要求"かを判定する。"定期負荷要求"でない時、S213で他の要求のチェックを行う。"定期負荷要求"であった時、S215でアプリケーション動作情報テーブル14をサーチし、バックアップサーバが定義されている動作アプリケーションの内、制御動作規約情報テーブルの同一サーバ構成時の固定バックアップ指定なしのアプリケーション(固定指示1109がOFFのアプリケーション)をサーチする。

【0052】該当するアプリケーションがサーチされると、S217でサーチされたアプリケーションのバックアップサーバの負荷と他サーバの負荷をサーバ負荷情報テーブルで比較しながら、より低負荷のサーバをサーチする。また、サーチしたより低負荷のサーバをサーバ状態情報テーブル12でサーチし、サーバ状態1203が正常であることを確認する。より低負荷で正常なサーバがあれば、S219で、該当アプリケーションのアプリケーション動作情報テーブル14の処理中フラグ1407を処理中にし、バックアップサーバ名1405をサーチしたサーバに書き換え、他サーバに反映する。次に、

S221でクラスタ管理機構900にバックアップサーバ名の変更指示を行う。完了後、アプリケーション動作情報テーブル14の処理中フラグ1407をリセットし、他サーバに反映する。(S223)。全アプリケーションに対し、上記S215～S223の処理を行い(S225)、処理が終了したらメッセージ待ちへ戻る(S227)。

【0053】9．制御動作例

以下に、制御動作例として、サーバ状態の変更に対するアプリケーションのバックアップサーバの割り付けの例をいくつか記述する。

(1)構成変更(異常停止／復旧)による動作例

5台構成時の構成変更時の割り付け例を示す。図20は、サーバ構成の変更例を示す図である。図21は、図20のサーバ構成の変更にそれぞれ対応するアプリケーションの割り付け例を示す図である。図20の20.01、図21の2101に示すように、各アプリケーションAP1，AP2，AP3がサーバ100a，100b，100cで動作し、各アプリケーションのバックアップサーバは、サーバ100dに割り当てられている。

【0054】次に、図20の2002に示すように、サーバ100aが異常停止すると、クラスタ管理機構900がアプリケーションAP1を、2101においてバックアップサーバに指定されているサーバ100dに移す。構成制御部40は、このトリガを元に、各アプリケーションのバックアップサーバをアプリケーションが何も動作していないサーバ100eに割り付ける。割り付けた状態を図21の2102に示す。

【0055】次に、図20の2003に示すように、サーバ100bが異常停止すると、クラスタ管理機構900がアプリケーションAP2をバックアップサーバに指定されているサーバ100eに移す。構成制御部40は、このトリガを元に、各アプリケーションのバックアップサーバをサーバ100c，100d，100dと、図21の2103に示すように再設定する。負荷による調整は、前述した説明のように行われるので、ここでは説明は省略する。

【0056】次に、図20の2004に示すように、サーバ100aが復旧して立ち上がり、それを構成制御部40が検知すると、このトリガを元に、構成制御部40は、各アプリケーションのバックアップサーバを、図21の2104に示すように、サーバ100aに再設定する。

【0057】(2)構成変更(復旧)による動作例

図22から図24は、異常縮退後の復旧時の動作例を示す図である。図22の2201、図23の2301に示すように、サーバ100d、100eが正常な状態で稼働しており、アプリケーションAP1、アプリケーションAP3がサーバ100dで重複動作しているものとする。その状態で、サーバ100cが復旧すると、構成制

御部4 0は、そのトリガを受け、サーバ1 0 0 c , 1 0 0 d , 1 0 0 e の3 台によるクラスタ構成時の制御動作規約情報テーブルのアプリケーションＡＰ３の自動移動モードを判定する。自動移動モードＯＮの場合、構成制御部4 0は、クラスタ管理機構9 0 0 に指示を出し、アプリケーションＡＰ３はサーバ1 0 0 c に移動(停止、再起動) される。その結果、図2 2 の2 2 0 2 、図2 3 の2 3 0 3 に示すように、サーバ1 0 0 c , 1 0 0 d , 1 0 0 e での運転に移行する。

【００５８】なお、自動移動モードＯＦＦの場合は、アプリケーションの移動は行われず、各アプリケーションのバックアップサーバにサーバ1 0 0 c が割り当てられる処理となる。この時の割り当てを、図2 4 に示す。

【００５９】(３) ＣＰＵ負荷による動作例

図2 5 から図2 8 は、ＣＰＵ負荷変動による定期更新の動作例を示す図である。図2 5 に示すように、サーバ1 0 0 c , 1 0 0 d , 1 0 0 e でアプリケーションＡＰ３, ＡＰ1 , ＡＰ2 がそれぞれ動作している。各ＣＰＵ負荷により図2 6 に示すように、バックアップサーバがサーバ1 0 0 c , 1 0 0 d , 1 0 0 d と割り付けられている。サーバ1 0 0 d で動作しているアプリケーションＡＰ1 のバックアップは、サーバ1 0 0 e より軽負荷であるサーバ1 0 0 c に割り付けられている。負荷更新部3 0 による負荷の更新後、図2 7 に示すように、サーバ負荷の軽負荷順位で、サーバ1 0 0 c とサーバ1 0 0 e が逆転したので、構成制御部4 0は、サーバ1 0 0 c をバックアップサーバとしていたアプリケーションＡＰ1 のバックアップサーバを図2 8 に示すように、サーバ1 0 0 e に切り替える。

【００６０】１０. 制御動作規約の定義

次に、制御動作規約の定義について、図2 9 を用いて説明する。共用データ記憶部1 0 の制御動作規約情報テーブル1 1 作成のために、ユーザが記述する定義ファイル2 9 0 1 について説明する。制御動作規約情報テーブル1 1 の作成方式は、規約記述型とモデル指定型の2 方式がある。

【００６１】ａ) 規約記述型

サーバ構成の各変動に対し、どのような動作制御を行うかの規約定義2 9 0 2 を詳細に記述したユーザ定義ファイル2 9 0 1 を図2 9 の2 9 0 3 に示すように、ユーザが記述して作成する方式である。このユーザ定義ファイル2 9 0 1 は、プログラミングを必要としない簡易なビルドインコマンドで記述可能とする。クラスタシステム立ち上げ時に構成制御部4 0により規約定義2 9 0 2 が参照され、制御動作規約情報テーブル1 1 が作成される。

【００６２】ｂ) モデル指定型

特定の動作モデルを予め定義し、その中からユーザに指定をさせる方式である。当然、内部で、指定されたモデルをもとにして、規約定義2 9 1 6 に変換が行われる。記述できるモデルは、以下の通りである。

・二重系ホットスタンバイ(ロードシェア含む) 型モデル

図3 0 に、二重系ホットスタンバイモデルの定義例を示す。このモデルでは、アプリケーションを双方のサーバで動くよう2 個用意した場合、ロードシェア型となる。アプリケーション動作中のサーバが異常停止した場合、動作中であったアプリケーションは、もう一方のサーバで再起動される。異常サーバが復旧した場合、復旧したサーバは、各アプリケーションのバックアップサーバとなる。

・多重系Ｎ対1 バックアップ(バックアップ浮動) 型

図3 1 に、多重系Ｎ対1 バックアップモデルの定義例を示す。(Ｎ＋1) 台のサーバで構成され、1 台を共通のバックアップとする形態である。アプリケーションはＮ個存在し、Ｎ台のサーバでそれぞれ動作する。動作サーバが停止すると、アプリケーションはバックアップサーバで動作する。異常サーバが復旧し、再度、組み込まれると、全アプリケーションのバックアップサーバとなる。なお、2 個以上のサーバが異常停止した場合、そのアプリケーションは切り捨てられ、サーバが復旧した時、再起動される。どちらの作成方式の場合も、記述ファイル内容はクラスタシステム立ち上げ時、図2 9 の2 9 0 5 , 2 9 1 8 に示すように、制御動作規約情報テーブル1 1 に展開される。

【００６３】以上のように、この実施の形態では、クラスタ(分散した複数の計算機を一まとまりにしたシステム) 構成に於いて、計算機及びアプリケーションの状態変化に応じて、アプリケーションの起動、停止、起動計算機の決定、及び引継ぎ計算機(アプリケーション、もしくはアプリケーション実行中の計算機が異常停止した場合、次にアプリケーションを実行する計算機のこと) の決定を自動的、かつ、最適に制御できるようにしたクラスタの自動構成制御方式について説明した。この実施の形態のクラスタシステムによれば、計算機及びアプリケーションの状態を常に監視し、その状態の変化に応じて、自動的かつ最適にアプリケーションの動作計算機を動的に制御するので、システム管理者の負担が軽減される。

【００６４】前述した実施の形態では、ネットワークを介して更新される共用データ記憶部をサーバ毎に備える場合について説明したが、クラスタシステム内で共用のメモリ装置を備えて共用データ記憶部としてもよい。

【００６５】

【発明の効果】この発明によれば、クラスタシステムを構成するサーバの状態に応じて、アプリケーションとサーバとの対応を動的に制御できる。これにより、システム管理者がシステム監視、制御に要する負担を大幅に軽減し、効率のよい高信頼システムの構築が可能になる。

【００６６】また、この発明によれば、クラスタに構成

されている計算機のメンバーに応じたアプリケーションの引継ぎ計算機を自動的に決定できる。

【００６７】また、この発明によれば、新たにサーバが起動した時、操作者を介入させずアプリケーションを自動引継ぎできる。

【００６８】また、クラスタのメンバーに変動が生じた際、実行中のアプリケーションの引継ぎ計算機を変更し、決定するので、常に最適な引継ぎ計算機を選択可能である。

【００６９】また、サーバの負荷に応じて、アプリケーションの引き継ぎを行うので、負荷の分散が図れる。

【００７０】また、定義ファイルを基にクラスタの構成を制御する規約情報を作成するので、プログラミングレスによる定義が可能である。

【００７１】また、システムで定義テンプレートを用意するので、パターン化された処理の場合、ユーザは、クラスタシステムの詳細な定義を行う必要がない。

【００７２】また、構成制御部が、定期的に取得された各計算機の負荷（ＣＰＵ、メモリ）状況によって最適な引継ぎ計算機を自動決定し、引継ぎを行うので、操作者が介入せず最適な引継ぎができる。

【図面の簡単な説明】

【図1】　この発明のクラスタシステムの構成を示すブロック図である。

【図2】　この発明のクラスタシステムの構成を示すブロック図である。

【図3】　この発明のクラスタシステムのハードウェア構成図である。

【図4】　この発明のクラスタシステムの機能を示すブロック図である。

【図5】　この発明のクラスタシステムの機能を示すブロック図である。

【図6】　この発明のクラスタシステムの動作概要を示す図である。

【図7】　この発明のクラスタシステムの状態監視部20が行うサーバ状態の監視を説明する図である。

【図8】　この発明のクラスタシステムの負荷更新部30が行う負荷状況の監視を説明する図である。

【図9】　この発明のクラスタシステムの共用データ記憶部１０の制御動作規約情報テーブル１１を示す図である。

【図１０】　この発明のクラスタシステムの共用データ記憶部１０のサーバ状態情報テーブル１２を示す図である。

【図１１】　この発明のクラスタシステムの共用データ記憶部１０のサーバ負荷情報テーブル１３を示す図である。

【図１２】　この発明のクラスタシステムの共用データ記憶部１０のアプリケーション動作情報テーブル１４を示す図である。

【図１３】　この発明のクラスタシステムの構成制御部４０の初期立ち上げ時の動作の流れ図である。

【図１４】　この発明のクラスタシステムの構成制御部４０のサーバ構成変更時の動作を示す流れ図である。

【図１５】　この発明のクラスタシステムの構成制御部４０のサーバ構成変更時の動作を示す流れ図である。

【図１６】　この発明のクラスタシステムの構成制御部４０のサーバ構成変更時の動作を示す流れ図である。

【図１７】　この発明のクラスタシステムの構成制御部４０のサーバ構成変更時の動作を示す流れ図である。

【図１８】　この発明のクラスタシステムの構成制御部４０のサーバ構成変更時の動作を示す流れ図である。

【図１９】　この発明のクラスタシステムの構成制御部４０の負荷更新時の動作の流れ図である。

【図２０】　この発明のクラスタシステムのサーバ構成の変更例を示す図である。

【図２１】　この発明のクラスタシステムの図２０のサーバ構成の変更に対応するアプリケーションの割り当て例を示す図である。

【図２２】　この発明のクラスタシステムの異常縮退後の復旧時の構成の変更例を示す図である。

【図２３】　この発明のクラスタシステムの図２２に対応するサーバの異常縮退後の復旧時のアプリケーションの割り当て例を示す図である。

【図２４】　この発明のクラスタシステムの図２２に対応するサーバの異常縮退後の復旧時のアプリケーションの割り当て例を示す図である。

【図２５】　この発明のクラスタシステムのＣＰＵ負荷変動による定期更新時の動作例を示す図である。

【図２６】　この発明のクラスタシステムのＣＰＵ負荷変動による定期更新時の動作例を示す図である。

【図２７】　この発明のクラスタシステムのＣＰＵ負荷変動による定期更新時の動作例を示す図である。

【図２８】　この発明のクラスタシステムのＣＰＵ負荷変動による定期更新時の動作例を示す図である。

【図２９】　この発明のクラスタシステムの共用データ記憶部１０の制御動作規約情報テーブル１１の作成を説明する図である。

【図３０】　この発明のクラスタシステムの二重系ホットスタンバイモデルの定義例を示す図である。

【図３１】　この発明のクラスタシステムの多重系Ｎ対１バックアップモデルの定義例を示す図である。

【符号の説明】

１０，１０ａ，１０ｂ，１０ｃ　共用データ記憶部、１１，１１ａ，１１ｂ，１１ｃ　制御動作規約情報テーブル、１２，１２ａ，１２ｂ，１２ｃ　サーバ状態情報テーブル、１３，１３ａ，１３ｂ，１３ｃ　サーバ負荷情報テーブル、１４，１４ａ，１４ｂ，１４ｃ　アプリケーション動作情報テーブル、２０　状態監視部、３０，３０ａ，３０ｂ，３０ｃ　負荷更新部、３５　制御動作

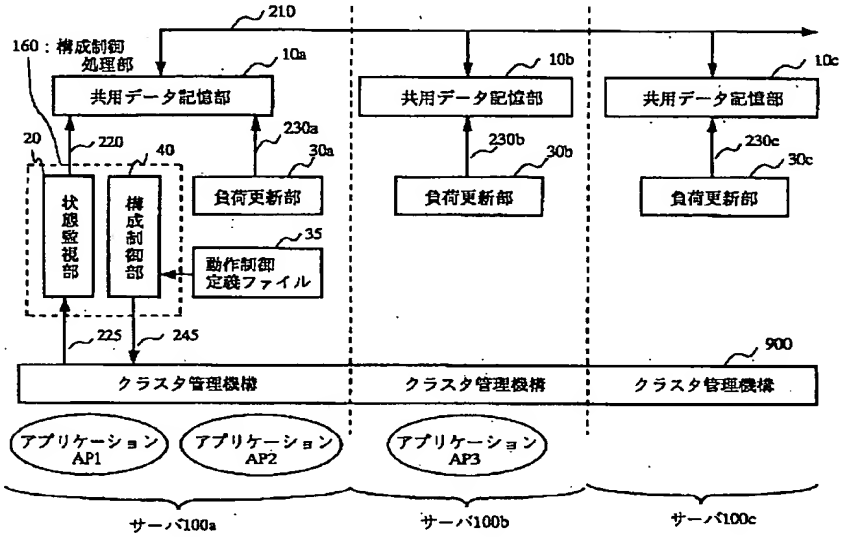
21

定義ファイル、４０　構成制御部、１００ａ，１００ｂ，１００ｃ，１００ｎ　サーバ、１６０　構成制御処理部、２００ａ，２００ｂ，２００ｎ　ディスク、９００クラスタ管理機構、９１０ａ，９１０ｂ，９１０ｃ　ＯＳツール群、１０００ネットワーク、１１０１　サーバ構成、１１０３　アプリケーション名、１１０５　動作サーバ、１１０７　バックアップサーバ、１１０９

22

固定指示、１１１１　自動移動モード、１２０１　サーバ名、１２０３　サーバ状態、１２０５検出状態、１２０７　処理モード、１３０１　サーバ名、１３０３　負荷情報、１４０１　アプリケーション名、１４０３　動作サーバ名、１４０５　バックアップサーバ名、１４０７　処理中フラグ、ＡＰ１，ＡＰ２，ＡＰ３　アプリケーション。

【図1】

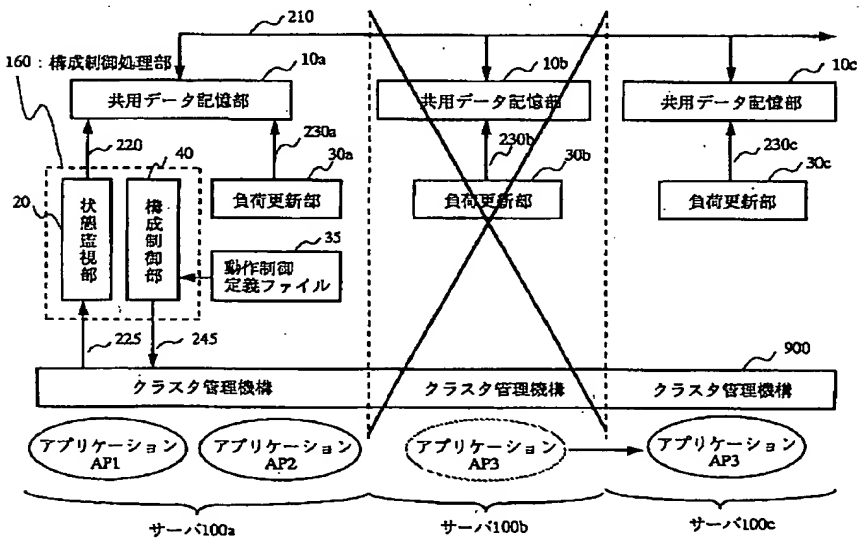


【図11】

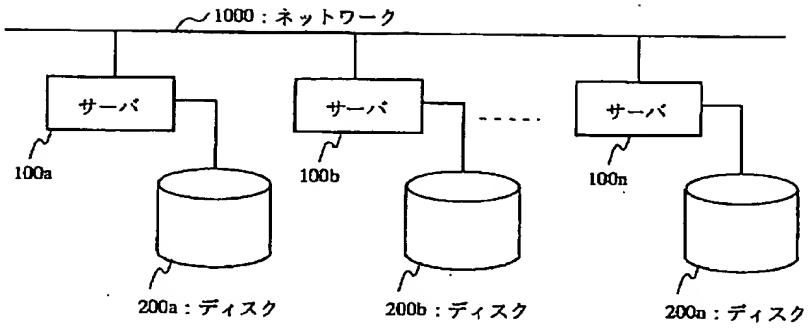
サーバ負荷情報テーブル　13

1301 サーバ名	1303 負荷情報
サーバ100a	0%
サーバ100b	40%
サーバ100c	60%
サーバ100d	20%

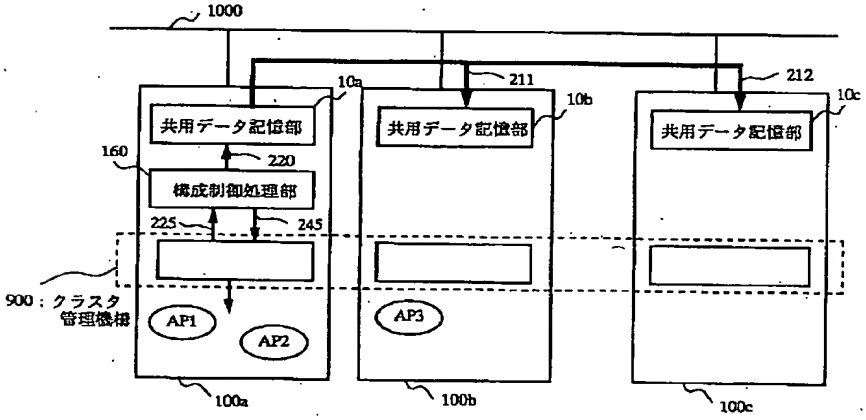
【図2】



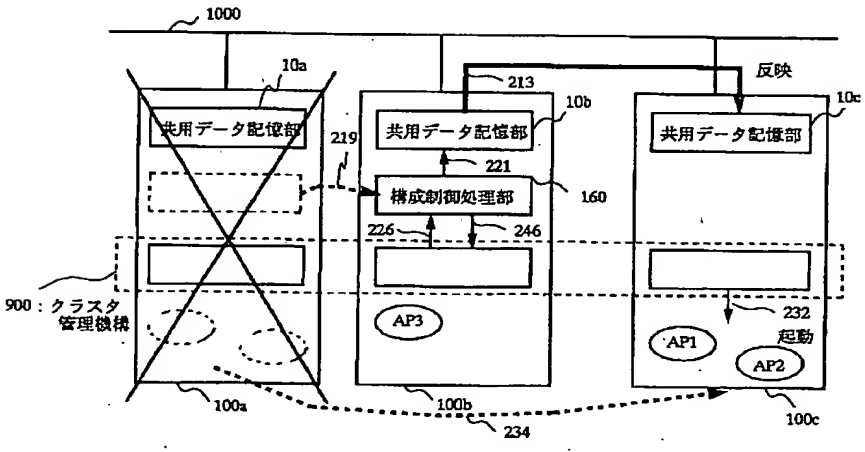
【 図3 】


1000：ネットワーク
サーバ
サーバ
サーバ
100a
100b
100n
200a：ディスク
200b：ディスク
200n：ディスク

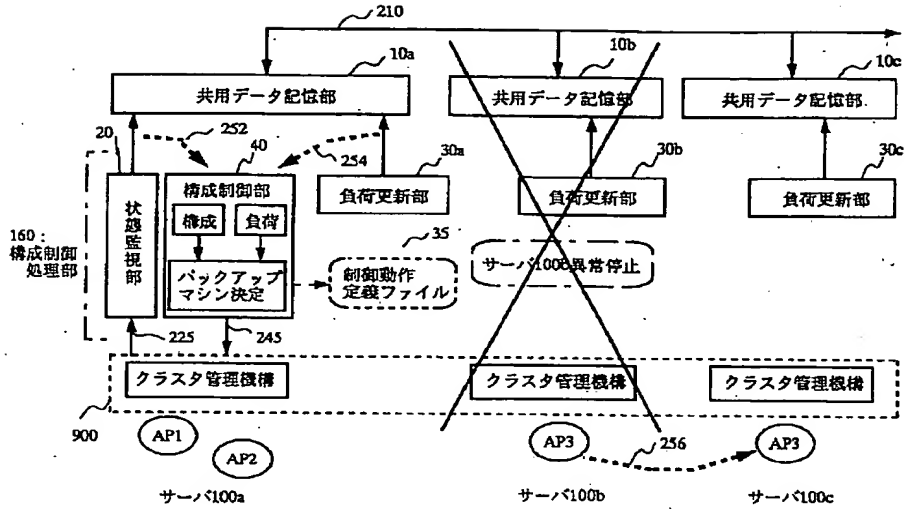
【 図4 】


1000
10a
211
212
10b
10c
共用データ記憶部
共用データ記憶部
共用データ記憶部
160
220
構成制御処理部
225
245
900：クラスタ
管理機構
AP1
AP2
AP3
100a
100b
100c

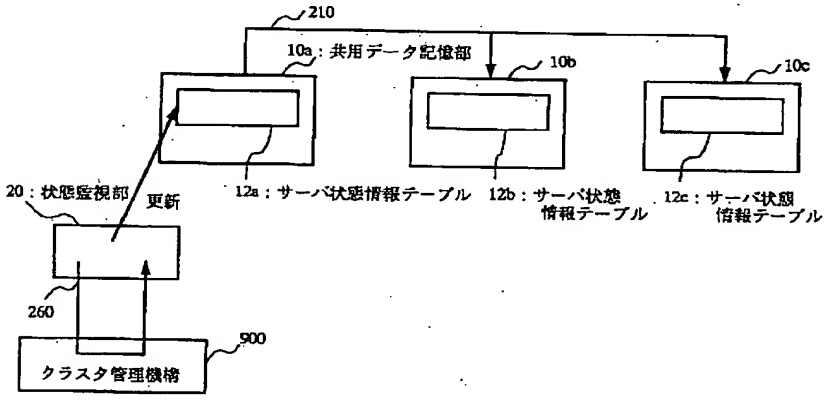
【 図5 】


1000
10a
213
反映
10b
10c
共用データ記憶部
219
共用データ記憶部
共用データ記憶部
221
構成制御処理部
160
226
246
900：クラスタ
管理機構
AP3
232
起動
AP1
AP2
100a
100b
100c
234

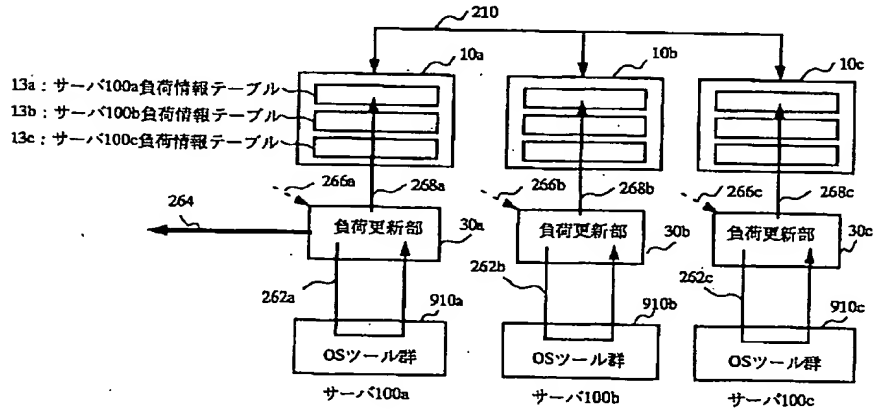
【 図6 】


210
10a
10b
10c
共用データ記憶部
共用データ記憶部
共用データ記憶部
20
252
40
254
30a
30b
30c
構成制御部
機成
負荷
状態監視部
負荷更新部
負荷更新部
負荷更新部
160：構成制御処理部
35
バックアップマシン決定
制御動作定義ファイル
サーバ100b異常停止
225
245
クラスタ管理機構
クラスタ管理機構
クラスタ管理機構
900
AP1
AP2
AP3
256
AP3
サーバ100a
サーバ100b
サーバ100c

【 図7 】


210
10a：共用データ記憶部
10b
10c
20：状態監視部
更新
12a：サーバ状態情報テーブル
12b：サーバ状態情報テーブル
12c：サーバ状態情報テーブル
260
900
クラスタ管理機構

【 図8 】


210
10a
10b
10c
13a：サーバ100a負荷情報テーブル
13b：サーバ100b負荷情報テーブル
13c：サーバ100c負荷情報テーブル
266a
268a
266b
268b
266c
268c
264
負荷更新部
30a
負荷更新部
30b
負荷更新部
30c
262a
262b
262c
910a
910b
910c
OSツール群
OSツール群
OSツール群
サーバ100a
サーバ100b
サーバ100c

【図９】

制御動作規約情報テーブル 11

サーバ構成 (1101)	AP名 (1103)	動作サーバ (1105)	バックアップサーバ (1107)	固定指示 (1109)	自動移動モード (1111)
サーバ100a、サーバ100b サーバ100c、サーバ100d	AP1	サーバ100a	サーバ100d	ON	OFF
	AP2	サーバ100b	サーバ100d	ON	OFF
	AP3	サーバ100c	サーバ100d	ON	OFF
サーバ100b、 サーバ100c、サーバ100d	AP1	サーバ100d	サーバ100c	OFF	ON
	AP2	サーバ100b	サーバ100d	OFF	ON
	AP3	サーバ100c	サーバ100d	OFF	ON
サーバ100c、サーバ100d	AP1	サーバ100d	サーバ100c	OFF	ON
	AP2	サーバ100d	サーバ100c	OFF	ON
	AP3	サーバ100c	サーバ100d	OFF	ON
			⋮		

【図１０】

サーバ状態情報テーブル 12

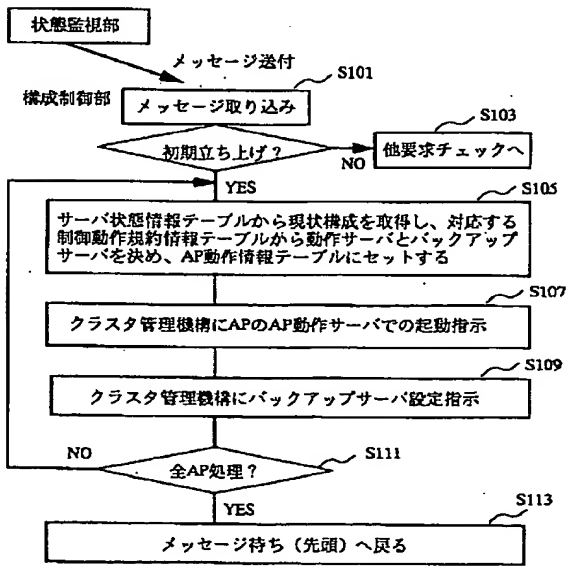
サーバ名 (1201)	サーバ状態 (1203)	検出状態 (1205)	処理モード (1207)
サーバ100a	切り離し		
サーバ100b	正常	異常停止	処理中
サーバ100c	正常		
サーバ100d	正常		

【図１２】

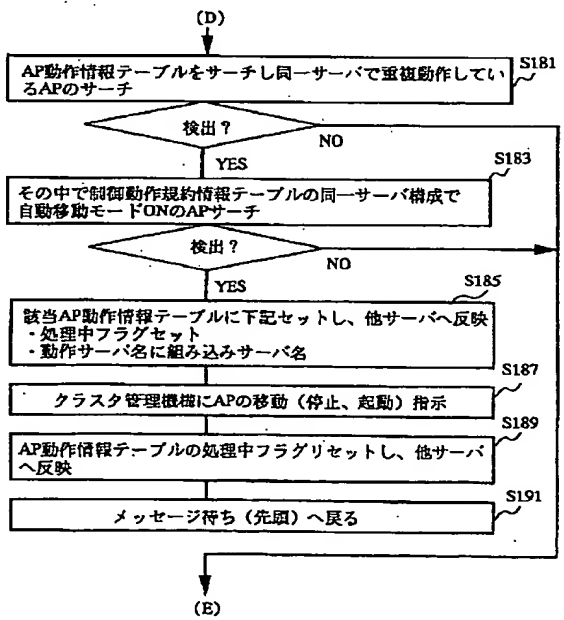
AP動作情報テーブル 14

AP名 (1401)	動作サーバ名 (1403)	バックアップサーバ名 (1405)	処理中フラグ (1407)
AP1	サーバ100d	サーバ100c	完了
AP2	サーバ100b	サーバ100d	処理中
AP3	サーバ100c	サーバ100d	

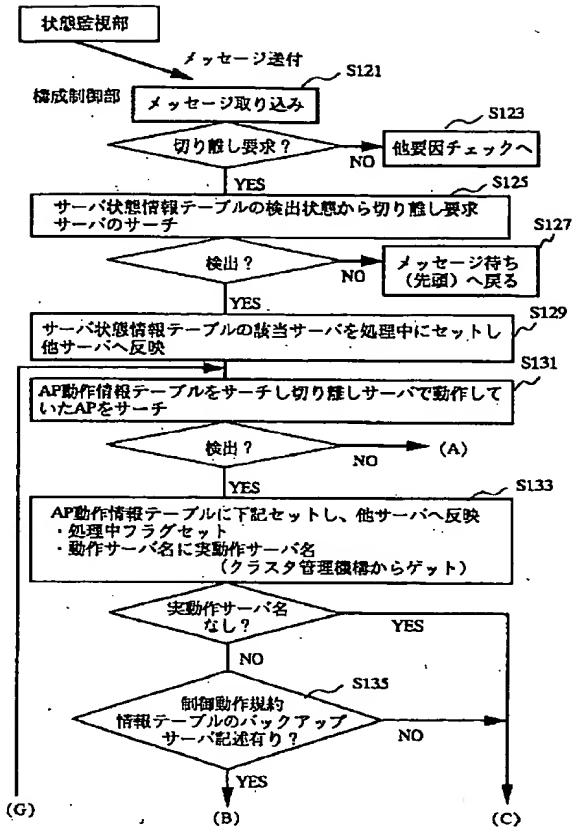
【図１３】



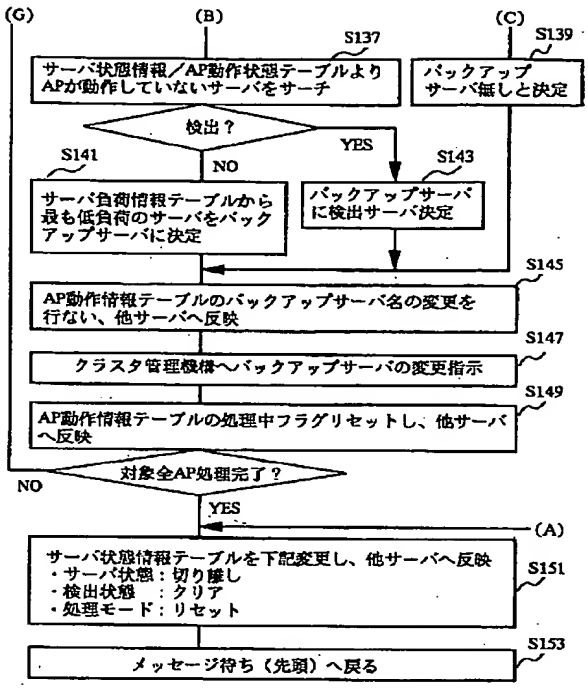
【図１７】



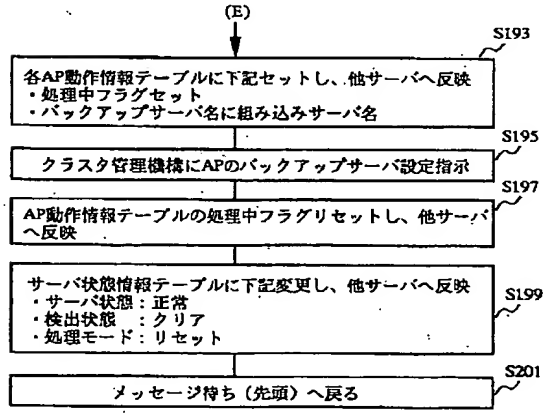
【 図１４ 】



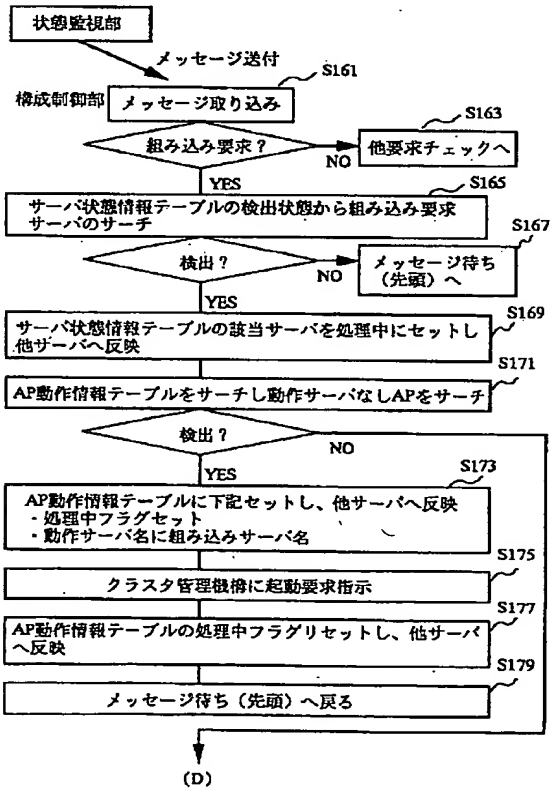
【 図１５ 】



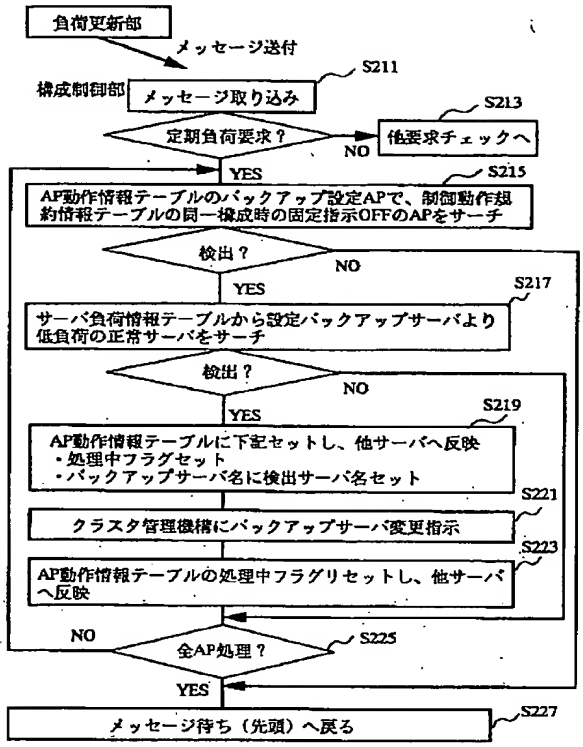
【 図１８ 】



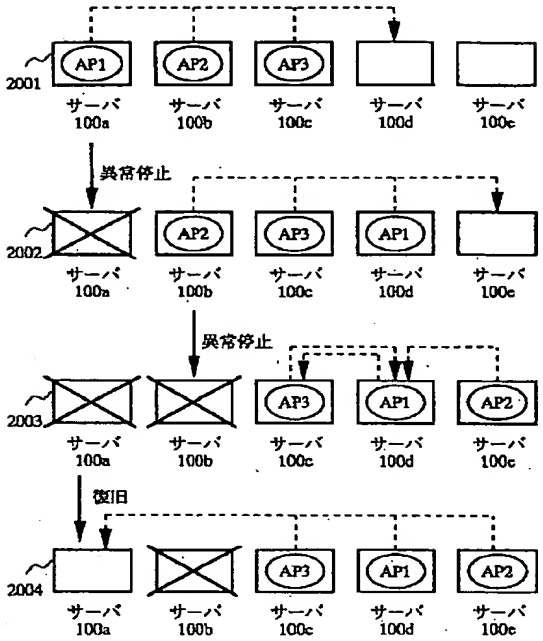
【 図１６ 】



【図１９】



【図２０】



【図２１】

AP名	動作サーバ	バックアップサーバ
AP1	サーバ100a	サーバ100d
AP2	サーバ100b	サーバ100d
AP3	サーバ100c	サーバ100d

2101

AP名	動作サーバ	バックアップサーバ
AP1	サーバ100d	サーバ100e
AP2	サーバ100b	サーバ100e
AP3	サーバ100c	サーバ100e

2102

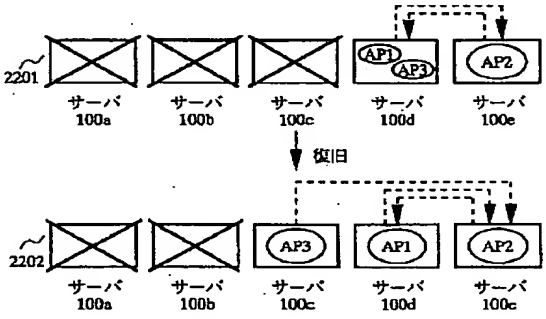
AP名	動作サーバ	バックアップサーバ
AP1	サーバ100d	サーバ100c
AP2	サーバ100e	サーバ100d
AP3	サーバ100c	サーバ100d

2103

AP名	動作サーバ	バックアップサーバ
AP1	サーバ100d	サーバ100a
AP2	サーバ100e	サーバ100a
AP3	サーバ100c	サーバ100a

2104

【図２２】



【図２４】

AP名	動作サーバ	バックアップサーバ
AP1	サーバ100d	サーバ100c
AP2	サーバ100e	サーバ100c
AP3	サーバ100d	サーバ100c

【図２３】

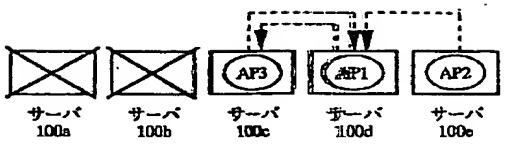
AP名	動作サーバ	バックアップサーバ
AP1	サーバ100d	サーバ100e
AP2	サーバ100e	サーバ100d
AP3	サーバ100d	サーバ100e

2301

AP名	動作サーバ	バックアップサーバ
AP1	サーバ100d	サーバ100e
AP2	サーバ100e	サーバ100d
AP3	サーバ100c	サーバ100e

2303

【図２５】

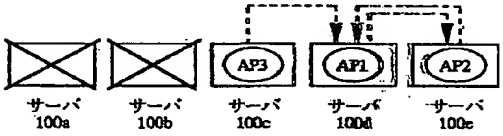


CPU負荷			40%	30%	50%
軽負荷順位			2	1	3

【図２６】

AP名	動作サーバ	バックアップサーバ
AP1	サーバ100d	サーバ100c
AP2	サーバ100e	サーバ100d
AP3	サーバ100c	サーバ100d

【図２７】

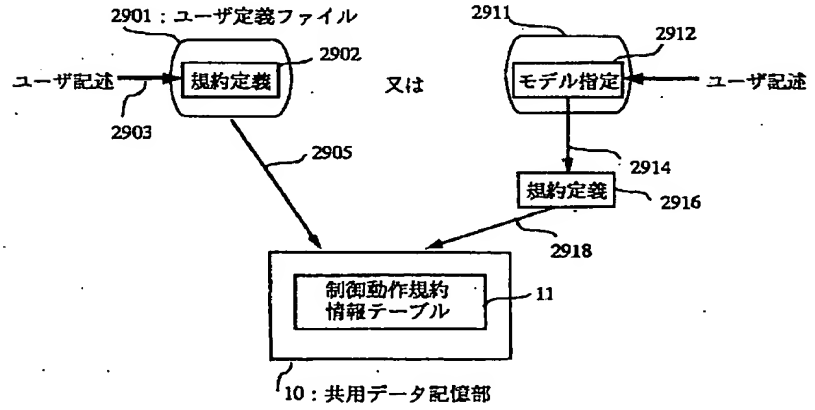


CPU負荷			60%	80%	50%
軽負荷順位			3	1	2

【図２８】

AP名	動作サーバ	バックアップサーバ
AP1	サーバ100d	サーバ100e
AP2	サーバ100e	サーバ100d
AP3	サーバ100c	サーバ100d

【図２９】



【 図３０ 】

モデル指定型：二重系ホットスタンバイモデルの定義ファイル

3001 サーバ構成	3003 AP名	3005 動作サーバ	3007 バックアップサーバ	3009 固定指示	3011 自動移動モード
サーバ100a、サーバ100b	AP1	サーバ100a	サーバ100b	OFF	ON
	AP2	サーバ100b	サーバ100a	OFF	ON

【 図３１ 】

モデル指定型：多重系N対1定義ファイル

3111 サーバ構成	3113 AP名	3115 動作サーバ	3117 バックアップサーバ	3119 固定指示	3121 自動移動モード
サーバ100a、サーバ100b、サーバ100c	AP1	サーバ100a	サーバ100c	OFF	ON
	AP2	サーバ100b	サーバ100c	OFF	ON